Merging Methodology Description, 14 Feb 12

The merging methodology is based on a hierarchy of sources. The merge begins with the source of highest priority (see table below). Then the station records in the source of the next lowest priority (candidate) is compared one by one against every other station in the higher priority source (master) to determine

(1) if the station record doesn't exist in the master dataset (i.e., is a unique station) and should thus be added in its entirety as a new station, or

(2) the station is present in both the master and the candidate source (and is thus a candidate for merging with the master source record).

a. Data from the candidate source will be added to the station record in the master source if it can add new data outside the period of record of the master station. Data will not be intermingled within a master's window of data. (e.g., a master station record having data from 1910 to 1950, a candidate station record having data from 1940 to 2010. Only the candidate data from 1950 to 2010 will be added to the master station record. No data within the period 1940 to 1950 will be added. This better ensures the homogeneity of the record.)

b. In some cases a candidate station record will add no new data and will not be merged (e.g., master source station record is 1910-1950, and candidate is 1925-1945. No merge will be made*.)

*An exception to this will be if the master station record has a lot of missing months and the candidate has much better temporal coverage – in this case the candidate could be used to overwrite the master in its entirety.

The determination of whether a candidate station should be merged or should be added as a unique station is based on 3 primary types of tests.

I. Metadata
II. Tests of overlapping data
III. Tests of non-overlapping data

I. Metadata tests

Three metadata characteristics are used to identify potentially matching or definitively unique stations.

(1.) Distance between stations

The distance between stations based upon latitude and longitude is fitted to an exponential decay function decaying from 1.0 at no distance to 0.0 at 100km and this value used as a probability that the two stations are the same.

(2.) Difference in elevation between the two stations.

The elevation difference also is fitted to an exponential decay function from 1.0 at no difference to 0.0 at 500 meters. This value is used as a probability that the two stations are the same.

(3.) Test of the similarity of the station name.

The Jaccard Index (JI), is used. This index is defined as the intersection divided by the union of two sample sets. The Jaccard Index, which has values from 1.0 to 0.0, looks for cases in which certain letters exist in both station names, as well as the number of times letters occur in one name, but not in the other.

These three geolocation metrics have a probability from 0 to 1. Using a simple Bayesian approach, they are multiplied using a weighted approach and a combined probability returned that the two stations are the same. Distance between stations has been shown to be the most dependable measure of similarity and is given a weight of 9. Station name the poorest so given a weight of 1, and elevation a weight of 5.

If this value surpasses a threshold of 0.5, the two stations are retained as possibly the same station and used for further evaluation by data comparisons. (This threshold is set low enough to account for the possibility that there are errors in metadata.) If the threshold is less than 0.5 for the candidate station, but 2 of the 3 metadata probabilities are greater than 0.9, the candidate station is held for further evaluation at a later time. Otherwise the candidate station is identified as a unique station and added to the higher priority source as a new station.

II. Data Comparisons – Overlapping Data

For a target and candidate station that has overlapping data, a direct comparison of observations during the same months and years is made. Three methods of evaluation have been preliminarily evaluated.

(a.) Normalized Root Mean Square Deviation (NRMSD)
(b.) Paired t-test
(c.) Index of Agreement (IA)

Preliminary evaluation of these measures showed that the paired t-test was ineffective at distinguishing unique from same stations. No further testing will be done.

Both the NRMSD and IA show promise for distinguishing unique from same stations. It is expected that two thresholds will be applied to either of these measures. A threshold above which there is great confidence that the candidate station is the same as the target station and should be merged. A threshold below which there is great confidence the candidate station is unique and should be added as new. Between the two thresholds there will be uncertainty and as such it will be necessary to place the candidate station and its data in a bin for closer examination at a later time. It is expected that there will be many stations classified as uncertain and not included in the first version of the merged Stage 3 dataset.

III. Data Comparisons – Non-overlapping data

If there are not at least 5 years on overlapping data, tests for non-overlapping data are applied. Tests have not yet been developed. Potential tests include a comparison of mean and standard deviation between the master and candidate non-overlapping periods.

Table 1. The 36 Stage-2 sources in the Databank and their priority as of January 24, 2012. These sources establish the foundation from which the Stage-3 merged dataset will be created.

Priority	Source	Priority	Source
1	GHCN-Daily raw (NCDC)	19	Spain (Univ. Rovira i Virgili)
2	Mexico (CDMP)	20	Russia (Roshydromet)
3	Vietnam (CDMP)	21	Uruguay (Inst. Nacional de Invest. Agropecuaria)
4	US Forts (CDMP)	22	Switzerland (Digihom/MeteoSwiss/IAC-ETH)
5	Channel Islands (States of Jersey Met)	23	Tunisia/Morocco (ISPD)
6	Ecuador (Inst. Nacional De Met E Hidrologia)	24	Europe/N. Africa (ECA Daily/KNMI)
7	Pitcairn Island (Met Service of New Zealand)	25	Southeast Asia (SACA/KNMI)
8	Beirut (Univ. of Giessen)	26	Japan (Japan Met Agency)
9	Brazil (INPE, Nat. Institute for Space Research)	27	UK Met Office Historical (UKMO)
10	Miscellaneous (NCDC)	28	Europe/N. Africa (ECA Monthly/KNMI)
11	World Weather Records (WMO)	29	GHCN-M v2 Source (NCDC)
12	Colonial Era Archives (Griffith)	30	GHCN-M v2 (NCDC)
13	East Africa (Univ. of Alabama-Huntsville)	31	Central Asia (NSIDC)
14	Antarctica South Pole (Univ. of Wisc.-Madison)	32	Canada (Env. Canada)
15	Switzerland (ISPD)	33	Australia (BOM)
16	Polar (ISPD)	34	Arctic (IARC/Univ. of Alaska Fairbanks)
17	Sydney (ISPD)	35	Greater Alpine Region (Histalp/ZAMG)
18	Antarctica (SCAR Reader Project)	36	HadCRUT3 (UKMO)